Appendix II Guidelines for corpus processing and the ripe corpus

This appendix comprises of two parts: Part I presents the guidelines for processing the corpus of verse lines: Section 1 briefly considers the analytical advantages coding offers as a means of corpus processing and Section 2 spells out the coding scheme. Part II presents the ripe corpus for each of the five genres which features the frequency pattern for each coding type.

Part I Guidelines for corpus processing

1 Coding as a means of corpus processing

The coding scheme to be proposed below represents the grammatical structure of the verse line by encoding the boundary strength between two surface adjacent syllables. This strength is attributable to the grammar, mostly syntax, occasionally supplemented by lexicon, semantic interpretation and pragmatic considerations. As such, coding may be regarded as an alternative to bracketing in representing the grammatical structure of the line. While bracketing suffices in the development of the modern verse grammar, coding is necessary in exploring the ancient grammar, because it offers considerably analytical convenience by better revealing the distinct patterns in the corpus that would be obscure otherwise. In particular, the numerical coding system greatly facilitates the distilling of the frequency patterns of lines of various grammatical structures and highlights the distribution patterns of both the weakest and the strongest boundaries, which is respectively coded as 1 and 4 below[1]. As shown in Chapter 7, both patterns constitute important evidence for the operativeness of the modern constraints in the ancient grammar.

In addition, coding offers certain additional advantages, two of which deserve brief mentioning. First, the coding scheme enables us to record the boundary strength without pinpointing the specific formal status of the syntactic constituents involved, which are often disparate. There is no fixed, one-to-one correspondence between the syntactic constituent and the boundary strength and constituents of disparate statuses may give rise to comparable boundary strength. This is illustrated in the following three verse lines where the syntactic bracketing and labels are given for illustrative purpose:

↓

(1) *[chu1ri4]*$_{\text{NP}}$*[[jing1 men2]*$_{\text{NP}}$ *shan1]*$_{\text{NP}}$
first sun Jingmen mountain
ëThe sun (rises on) the Jingmen mountainí

[1] In theory, it is also possible to just use bracketing to obtain such patterns from the corpus. However, the difficulty in reading and keeping track of brackets would result in the corpus to be processed in a much more cumbersome and less efficient fashion.

(2)
↓
[mu4 luo4]$_{S}$ *[yan4 [nan2du4]]*$_{S}$
tree fall swan south cross
ëThe tree leaves fall and the swans fly to the southí

↓
(3) *[xing4]*$_{ADJUNCT}$ *[[yan3 ming2]*$_{S}$*[shen1 jian4]*$_{S}$*]*
luckily eye bright body healthy
ëLuckily I am still bright-eyed and in good healthí

The three boundaries marked out with the arrows are equally strong in that they all represent the strongest structural boundaries within the line and as such will be uniformly encoded as 4 in our coding scheme (to be introduced below). However, as indicated by the syntactic labels, the syntactic constituents involved differ considerably in nature.

Second, in a related manner, a numerical coding scheme caters to the relative nature of boundary strength. The domain of the coding is limited to the verse line, and the strength of a boundary between two adjacent syllables in a line is always gauged in relation to that of the other boundaries in the same line. Across the lines, syntactic constituents of different statuses might trigger the same boundary strength; conversely, syntactic constituents of the same status might feature different boundary strengths. But the numerical coding scheme enables boundaries of different origins to be represented uniformly.

As a final note, we wish to emphasize that coding is, in essence, a notational shorthand which offers the above-mentioned analytical convenience but carries no theoretical import. It can be translated into bracketing, although they are not in a one-to-one relation. Several coding types may correspond to the same bracketing structure, as shown in Chapter 7.

2 The coding scheme

This section presents the coding scheme which is a numerical way to encode the boundary strength by uncovering and incorporating the linguistic factors responsible for this strength. Evidently, linear adjacency of two syllables is the premise for the following discussion on the boundary strength and the coding scheme.

The scheme is five-scaled with the numbers ranging from 1 to 5, where 1 indicates the weakest boundary and 5 the strongest. The smaller the number, the weaker the boundary[2]. This is indicated below:

(4) weakest boundary <-----1-----2-----3-----4-----5-----> strongest boundary

The coding process contains three steps: (i) pre-assignment, (ii) assignment and (iii) post-assignment, which are respectively discussed below.

[2] Five scales are chosen in an effort to achieve a balance between descriptive sufficiency and analytical efficiency (cf. Chen 2000: 563; Hayes 2000).

2.1 Pre-assignment

The purpose of pre-assignment is to encode those boundaries whose strength can be straightforwardly determined in order to clear the road for the more elaborate assignment stage which takes recourse to syntax and lexicon. The codings 1, 3, and 5 are assigned at this stage.

First, coding 1 is assigned to weakest boundaries which, in the context of classical Chinese verse, include the boundary (i) between the reduplication and disyllabic morphemes, and (ii) between the component syllables in opaque proper names, i.e. place and person names[3]. In both cases, nothing can be inserted in between and the two syllables cannot be split in scrambling. Reduplication is quite common in classical Chinese and used widely in verse, typically in onomatopoeic words such as ë*xiao1 xiao1*í (the sound of falling leaves) or adjectives reduplicated for more vivid effect, e.g. ë*qing1 qing1*í (green). Disyllabic morphemes are relatively rare due to the overwhelmingly monosyllabic nature of classical Chinese; some typical examples are the names of flora and fauna, for instance, ë*pi4 li2*í (a kind of plant) and ë*ju1 jiu1*í (a kind of seabird).

Second, coding 5 is, rather trivially, assigned to, and only to the boundary following the line-final syllable, for the simple reason that the end of the line, with no syllable to follow, vacuously qualifies as the strongest boundary in the line. A noteworthy point here is that a verse line may correspond to a wide array of syntactic constituents such as a phrase, a phrase coordination, a sentence, a compound sentence consisting of two or more small clauses. These are respectively illustrated below:

(5) *shi2 nian2 li2 luan4 hou4*
ten year separation chaos after
ëAfter ten yearsí separation and chaosí

(6) *gu3 dao4 xi1 feng1 shou4 ma2*
ancient road west wind thin horse
ëThe ancient road, the west wind, and the thin horseí

(7) *gu4 ren2 ju4 ji1 shu3*
old folks prepare chicken rice
ëThe old friends have prepared chicken and riceí

(8) *zhu3 xuan1 gui1 huan4 nu3*
bamboo noisy return washing girl
ëThe bamboo (leaves) become noisy, (and) the washing girl returnsí

Finally, coding 3 is assigned to all the remaining boundaries, but only temporarily as a default coding to be modified below.

[3] By ëopaqueí we refer to those proper names whose meaning cannot be compositionally derived, e.g. ë*zhu1 ge3*í (person name) or ë*yue4 yang2*í (place name). This is in contrast to those ëtransparentí proper nouns where the meaning can be so derived, for example, place names such as ë*jing1 zhou1*í where ë*zhou*í means ëcityí, and personís names such as ë*wang2 gong1*í where ë*gong*í means ëlordí. Such transparent proper names are in fact compounds or NPís, which, as is to be argued shortly, feature binding factors.

2.2 Algorithm for the assignment

This phase of the coding process is targeted at the boundaries temporarily assigned coding 3 in the pre-assignment phase, which entails a close scrutiny of the grammatical structure of the verse line. The algorithm features ëbinding factorsí and ëalienating factorsí: at those boundaries where binding factors are present, the current coding (which is the default 3) is reduced by one, thus becoming 2, whilst at boundaries where alienating factors are present, the coding 3 is increased by one, thus becoming 4. Below the binding and alienating factors are respectively spelled out.

2.2.1 The binding factors

Three types of binding factors can be identified: (i) certain semantic relations encompassed in the argument structure; (ii) inclusion in the lexicon; (iii) cliticization.

2.2.1.1 Semantic relations in the argument structure

The construct of argument structure (Williams 1981, 1994) is adopted to capture the relevance of syntactic structures to boundary strength. Briefly speaking, the argument structure of a lexical item, typically a predicate, is the lexical representation of its grammatical information (Grimshaw 1990). A distinction is drawn between internal and external arguments of a predicate in terms of whether an argument appears within the maximal projection of a predicate or not.

Two semantic relations, namely, the theta relation and functor relation, constitute the first binding factor. First, the theta relation refers to the syntactic relation between the predicate and its argument(s). In particular, the juncture between the predicate and its internal argument, most typically, that between a verb and its object NP, is the ëtightest of all grammatical relationsí, and is ëessentially as tight as it can getí (Williams 1994:29)[4]. In other words, such boundaries are the weakest. So is the boundary between a preposition and its complement NP in a PP. Examples are the VPís ë*ba3 jiu3*í and ë*wen4 qing1 tian1*í and the PP ë*sui2 chun1*í in the two verse lines below. The boundaries involving the theta relation are marked out.

↓ ↓
(9) *[ba3 jiu3] [wen4 qing1 tian1]*
hold wine ask blue sky
ëHolding the wine, (I) ask the blue skyí

↓
(10) *hu2die2 bu4* ***[sui2 chun1]*** *qu4*
butterfly not with spring leave
ëThe butterflies do not leave with the springí

In this connection, the boundary between the predicate and its external argument, typically that between the verb and its subject NP, is not characterized by the binding factor[5]. This is because unlike the internal argument which is in an immediate sisterhood relation with the verb, the external argument lies external to the maximal

[4] One should be careful not to confuse the use of ëjunctureí in Williams (1994) with that of ëboundaryí here: a tight juncture is a weak boundary.

[5] It does not constitute an alienating factor either; the coding at such boundaries retains the default ë3í, subject to promotion to 4 in the post-assignment stage.

projection of the verb, and the theta role is only assigned via the x-bar projection. As such, the external argument is not strictly ëlocalí to the verb. Indeed, according to Williams (1994:21), the subject-predicate juncture is a double-headed, phrase-to-phrase link, in contrast to the verb-object juncture which is single-headed and lexical. The relatively strong boundary between the verb and its external argument compared to that between the verb and its internal argument is evident from the much greater mobility enjoyed by the subject NP than the object NP, which often brings the subject NP out of linear adjacency with the predicate.

The second relation in the argument structure theory that serves as a binding factor is the functor relation. It differs from the theta relation in that it is neutral regarding theta roles. However, it is similar to the theta relation, or more precisely, the relation between the predicate and its internal argument in that both observe absolute locality and nothing can be inserted in between. Williams (1994:45) presents an inventory of constructions characterized by the functor relation, which are essentially reducible to the ëmodifier + modifieeí type. Among them, the relevant ones in the current context of classical Chinese verse are: (i) modifier + noun; (ii) (verbal) adverb + verb; (iii) negation + VP/AP.

First, in the ëmodifier + nouní construction, the modifier is either an adjective or noun. In both cases, the boundary between the modifier and the modifiee, i.e. the head noun, is weak. For example,

(11) ***[shen1↓yuan4]*** *suo3* ***[qing1↓qiu1]***
deep yard lock lonely autumn
ëThe lonely autumn is locked inside the deep yardí

(12) ***[chun1↓hua1]*** ***[qiu1↓yue4]*** *he2 shi2 liao3*
spring flower autumn moon whichtime disappear
ëWhen will the spring flowers and autumn moons disappear?í

which respectively contain NPís of the structure A+N and N+N, and where the relevant boundaries marked out are all weak.

A further piece of evidence for the weak boundary in the ëmodifier + nouní structure is the strong tendency to lexicalization displayed by such structures. Indeed, Duanmu (1998, 1999) argues that such structures are all compounds rather than noun phrases in modern Chinese. A similar picture is presented for such structures in classical Chinese in Feng (1998), which suggests that in classical Chinese, A/N+N structures were most likely to undergo idiomatization and become lexicalized into nominal compounds, especially when they were used with considerable frequency[6].

[6] Apparently, this bears on the issue of the distinction between noun phrases and nominal compounds, which, albeit interesting, is of little immediate relevance to the present discussion of boundary strength, since whether a given A/N + N structure is phrasal or lexical, the ëbinding factorí, being either the functor relation or the listed entry in the lexicon, is always present and thus the boundary between the two components is always weak. We will return to this issue when discussing the factor of inclusion in the lexicon below.

The second construction featuring the functor relation is ë(verbal) adverb + verbí. The ëverbalí adverb, which modifies the VP, is distinguished from the ësententialí adverb, which modifies the sentence. This is illustrated below with the sentential adverb ë*xing4*í and the verbal one ë*du2*í:

(13) ***xing4*** *[[yan3 ming2]*$_{S}$*[shen1 jian4]*$_{S}$*]*
luckily eye bright body healthy
ëLuckily I am still bright-eyed and in good healthí

(14) *lou2 shang4 hua1 zhi1 xiao4* ***[du2*** *mian2]*$_{VP}$
boudoir above flower branch laugh lone sleep
ëThe girl upstairs in the boudoir laughs at me sleeping aloneí

In terms of semantic relation, both subcategories of adverbs entertain a functor relation with their modifiees. However, only the ëverbal Adverb + VPí construction, as that in (14), contains the binding factor and accordingly the internal boundary is weak. The reason is that the modifiee of the sentential adverb, i.e. the sentence (IP), occupies a structurally higher node than VP; in other words, the sentence has a more elaborate branching structure than VP. As is to be seen in the next section, branching constitutes an alienating factor, which strengthens the boundary between the sentential adverb and the sentence it modifies. Indeed, the boundary between the sentential adverb and the sentence it modifies is typically the biggest break in a line and coded as 4. The cancellation effect between the binding and the alienating factors in the case of sentential adverbs renders the boundary between a sentential adverb and the modified sentence stronger than that between a verbal adverb and the modified verb.

It deserves mentioning here that similar to the ëA/N + NPí structure, the ë(verbal) Adverb + VPí structure is also susceptible to lexicalization, which further indicates the close tie between the adverb and the verb[7].

Third, the negation construction is another type of the ëmodifier + modifieeí structure, with the modifier being the negator ë*bu4*í and ë*wei4*í (meaning ënotí) and the modifiee typically being VP or AP, as shown in the following examples:

(15) (i) *meng4 jun1 jun1* ↓***bu4****zhi1*
dream you you not know
ëI dream of you, but you do not knowí

(ii) *heng2 zhi1* ***wei4*** ↓ *ye4*
horizontal branch not leave
ëThe horizontal branches have not yet grown leavesí.

[7] It needs to be realized, however, that there are far fewer verbal compounds deriving from the latter structure due to the smaller number of verbal adverbs. Some examples are ëshen4 si1í (carefully consider) and ëchang2 tan4í (give a long sigh over).

(16)

↓

(i) *feng1* **bu4** *ding4*
wind not certain
ëThe (direction of the) wind is not certainí

↓

(ii) *hong2 yan2* ***wei4*** *lao3 en1 xian1 duan4*
red complexion not old favor first stop
ëThe beauty is not yet old, but (the emperor) already loses favor of herí

The functor relation in such constructions renders the boundary between the negator and the following VP/AP weak. In fact, the weak boundary can also be accounted for by treating the negator as ëa lexical item not specified for categoryí, following Williams (1994:49), rather than as an adverb. This way, ë*bu4*í is a head that takes what it modifies as the complement constituting a so-called ëNotPí, and the modifiee serves as a NotP internal argument. If this account holds, then the binding factor between the negator and what it negates is attributable to a relation equivalent to that between the predicate and its internal argument. Whichever option is taken, the boundary in the negation construction is weak.

A further indication of the weak boundary between the negator and the constituent it negates is that the negation construction ë*bu4* + VP/APí is also susceptible to lexicalization, in particular when the VP/AP only comprises a monosyllabic verb or adjective, e.g. ë*bu4 duo1*í (not much/many), ë*bu4 xiang3*í (not want), and ë*bu4 zhi1*í (not know) in (15).

The range of syntactic construction types covered in the above discussion about the first binding factor, namely, theta and functor relations, actually encompass the majority of syntactic structures in classical Chinese verse lines, which are distinctly characterized by a minimal use of function words[8]. The following table summarizes these constructions and their respective semantic relations. Due to the presence of the binding factor, the boundaries in such constructions are all weak. The two semantic relations are respectively shortened as ëthetaí and ëfunctorí. In the case of the negation construction, corresponding to the two viable accounts mentioned above, both the functor and the theta relations are presented as the possible semantic relation.

(17)

Syntactic construction Type	**Binding factor**	**Boundary**	**Semantic relation**
V + NP	Yes	Weak	Theta
P + NP	Yes	Weak	Theta
A/N + NP	Yes	Weak	Functor
Verbal Adv+VP	Yes	Weak	Functor
Negator + VP/AP	Yes	Weak	Functor /theta

[8] Indeed, two of the five genres exclusively use lexical categories, and as is to be seen below, in the other three genres, only a very small number of function words are used.

2.2.1.2 Inclusion in the lexicon

The second binding factor is lexical in nature: a compound that is listed in the lexicon has a binding factor between its component syllables[9]. Such compounds could be nominal, verbal or adjectival, and their internal structures could be coordination or subordination (i.e. modification). Some examples of compounds are given below and the relevant boundaries are marked out.

(18) N+N coordination:

↓
[feng1 yu3] *rao4 cheng2 ai1*
wind rain surround city sad
ëThe wind and rain surround the city sadlyí

(19) N+N modification

↓ ↓
[chun1 hua1] [qiu1 yue4] *he2 shi2 liao3*
spring flower autumn moon which time disappear
ëWhen will the spring flowers and autumn moons disappear?í

(20) A+N modification; N+N coordination

↓ ↓
qi1 qi1 ***[fang1 cao3]*** *yi4* ***[wang2 sun1]***
luxurious fragrant grass miss kings lords
ëThe fragrant grass is so luxurious, and I am missing the kings and lordsí

(21) N+N coordination; A+A coordination

↓ ↓
[shen2 hun2] [mi2 luan4]
spirit spirit confused chaotic
ëThe spirits are confused and chaoticí

(22) A+N modification; V+V coordination

↓ ↓
[yu4 jie1] *kong1* ***[chu4 li4]***
jade stairs futile stand stand
ë(I) futilely stand on the jade stairsí

(23) verbal Adverb+V modification

↓
[xie2 yi3] *xun1 long2 zuo4 dao4 ming2*
obliquely lean fragrant pillow sit till dawn
ë(She) obliquely leans against the fragrant pillows and sit (in bed) till dawní

Compare the compounds of the modification type presented here with the ëmodifier + modifieeí constructions discussed earlier and the borderline between compounds and phrases in such cases seems blurry. This is especially true of the boundary between disyllabic NP or VP and disyllabic noun or verb compounds, as widely acknowledged among Chinese linguists (cf. Feng (1998) for classical Chinese and Duanmu (1998) for modern Chinese). In most cases, the crux seems largely a matter of frequency of usage: according to Feng (Ibid.), compounds may be regarded as idiomatized phrases, i.e., phrases that have become lexicalized due to their high frequency of usage.

[9] Actually, this argument has already been exploited in the above discussion when we cited the proneness for lexicalization as an indication of a weaker boundary.

However, this ambiguity has no bearing on the boundary strength under discussion here: whether a ëN+Ní, ëA+Ní, or ëAd+Ví structure constitutes a phrase or compound, the boundary between the two adjacent syllables involved is weak[10].

2.2.1.3 Cliticization

As mentioned earlier, classical Chinese verse is characterized by the parsimony, and in some genres, absence, of function words. With one exception, all the boundaries involving the few function words that do occur can be accounted for via the two binding factors discussed so far. This exception is the boundary involving the function word ë*zhi1*í in three usages, i.e. as the possessive marker, the particle linking subject and predicate, and the demonstrative pronoun, as respectively illustrated below:

↓
(24) (i) *gao1 yang2* ***zhi1*** *pi2*
lamb sheep ís skin
ëThe skin of the lambs and sheepí

↓
(ii) *zhi2 zi3* ***zhi1*** *shou3*
hold you ís hand
ë(I) hold your handí

↓
(25) (i) *han4* ***zhi1*** *guang3 yi3*
han (state name) particle wide particle
ëThe state of Han is wideí

↓
(ii) zi3 ***zhi1*** *bu4 shu1*
you prt not nice
ëYou are not niceí

(26) ↓
zhi1 *zi3 yu2 gui1*
this person go return
ëThis person is goingí

In all usages, ë*zhi1*í serves as a proclitic (Chen 1996: 598), and its rightward cliticization constitutes a strong binding factor between ë*zhi1*í and its following syllables.

Although these three usages of ë*zhi1*í are the only cases of cliticization as a binding factor, the above discussion prompts us to quickly examine one further usage of ë*zhi1*í and the other function words, which has so far remained undiscussed.

First, in addition to the above-mentioned three usages, ë*zhi1*í can also be used as the object pronoun, as shown below:

[10] One might also argue that the binding factor, being essentially a semantic relation (functor relation), is to some extent independent of the grammatical status of the structure.

(27) ↓
qiu2 ***zhi1*** *bu4 de2*
desire her no obtain
ë(I) desire her, but (I) cannot get (her)í

ë*Zhi1*í in this usage behaves like a full noun and the boundary between it and the preceding verb in this usage is that between the verb and its internal argument and thus weak.

The other function words occurring in our corpus are the possessive pronoun ë*qi2*í, the conjunction ë*qie3*í (and) and ë*er3*í (and), and several interjections. They are respectively illustrated below:

(28) *dai4* ***qi2*** *ji2* *xi1*
wait his kindness interj
ëAh, (I) wait for his kindnessí

(29) *xun2* *mei3* ***qie4*** *yi4*
bright beautiful and different
ë(She is) so bright, beautiful and differentí

(30) *xin1* ***er3*** *chang2* *xi1*
slim and long interj
ëAh, (he is) slim and tallí

The boundary between the possessive pronoun ë*qi2*í and its following N in (28) is comparable to the A+N structure and thus weak. The conjunction in (29) and (30) heads a constituent like the ëandPí in English (Williams 1994:16), which is similar to the negation structure in that the constituent following the conjunction serves as its internal argument, and accordingly the boundary is weak.

By comparison, interjections constitute an alienating factor, which will be discussed in the next section.

To sum up, three binding factors are identified: first, the two semantic relations encompassed in the argument structure theory, i.e. the theta relation and the functor relation; second, the inclusion as a lexical entry; third, cliticization. In terms of coding, the presence of any one of these binding factors at a boundary triggers the boundary strength to be reduced by 1.

2.2.2 The alienating factors

Two alienating factors are identified: branchingness of a structure and presence of interjections. Regarding the former, two points merits attention[11]. First, we stipulate that the coding of a boundary is only increased by one no matter whether the structure branches on one or both sides of it. Second, the alienating and the binding factors work independently of each other. For example, in a ëVerb + object NPí structure where the NP branches, the boundary between V and NP features both a binding factor and an alienating one, respectively due to the theta relation and the

[11] We assume the relevance of branchingness in syntax, as is evident from the order of verb clusters (Haegeman and van Riemsdijk 1986) and c-command.

branchingness of the internal argument NP. This is illustrated by the boundary between the verb ëtou4í and its complement NP ëbo2 luo2 shang3í below:

↓

(31) *ye4 han2 wei1 tou4 [bo2 [luo2 shang3]]*
night chill slightly penetrate thin gauze skirt
ëThe night chill slightly penetrates her thin gauze skirtí

Thus, the coding at this boundary first moves from the default 3 (assigned at the pre-assignment stage) to 2 (=3-1) due to the theta relation, and then is increased by 1 due to the branchingness, thus eventually arriving at 3 (=2+1).

A second alienating factor is the interjection: we contend that interjections, which are, by their very nature, semantically empty and syntactically unattached, stand in a loose relationship with their surrounding syllables. Accordingly the boundary between an interjection and its neighbors is strong and its coding is increased by one. More specifically, the boundary before a line-final interjection always constitutes the biggest break in the line, while a line-medial interjection triggers strong boundaries on both of its sides. In our corpus, there is only one line-medial interjection, i.e. ëxi1í, and the two boundaries bordering it are both strong, as shown below:

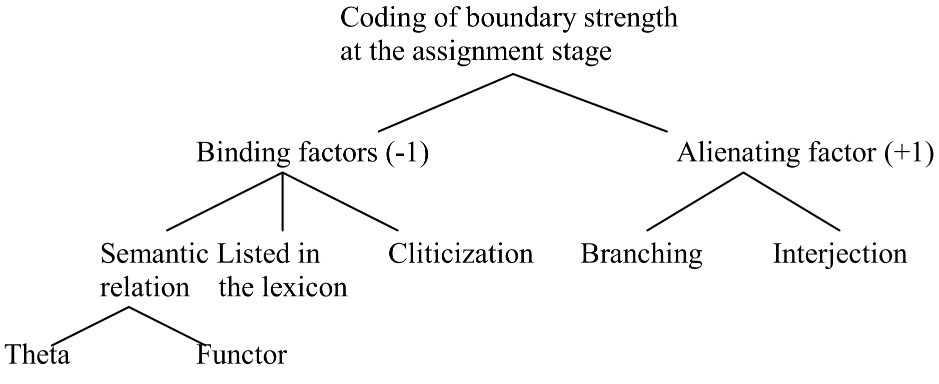
↓ ↓

(32) *[jia4 fei1 long2] xi1 [bei3 zheng1]*
ride fly dragon xi north march
ë(I) ride the flying dragon and go to the northí

2.2.3 Overview of the algorithm for coding assignment

Below is an overview of the algorithm for encoding boundary strength at the assignment stage:

(33)



2.3 Post-assignment

If the assignment stage is mainly concerned with the local addition or deduction of boundary strength coding, the post-assignment stage examines the well-formedness of the global coding profile emerging out of the pre-assignment and the assignment stages and straightens up possible irregularities. This entails a top-down perspective which differs from the bottom-up one at the assignment stage.

The following coding profile template is assumed for every verse line: one and only one 5 which necessarily occurs line-finally, one and only one 4 which encodes the strongest boundary within the line (except for two cases to be discussed below), zero or more 1ís, zero or more 2ís and zero or more 3ís. The coding profile reached at the end of the assignment stage is compared against this template and when the template fails to be met, adjustments are made accordingly.

Specifically, adjustments are called for when (i) more than one 5 is assigned, and/or (ii) no 4 or more than one 4 is assigned. Such cases could arise as a result of the implementation of the algorithm at the assignment stage, which is in turn based on the coding reached at the pre-assignment stage. First, of the multiple 5ís, only one is assigned at the pre-assignment stage and all the others are derived at the assignment stage from the default coding 3. We refer to these 5ís respectively as ëunderivedí and ëderivedí. Given the template outlined above, all the derived 5ís are demoted into 4ís. Second, we stipulate that there is only one 4 in the coding template which marks the strongest boundary in the line. As a consequence, when the coding at the end of the pre-assignment and the assignment stages contains no 4ís, the 3 at the strongest boundary in the line is promoted into 4; when the coding contains multiple 4ís, all the others except the one at the strongest boundary are demoted into 3ís[12]. Evidently, the post-assignment stage is not as trivial as the pre-assignment stage since it involves the determination of the strongest boundary in the line, which is discussed below.

2.3.1 Coding 4 at the strongest boundary in the line

In most cases, the strongest boundary in the line can be determined on syntactic grounds, although as in the case of the coding 5 boundary, the coding 4 boundary may correspond to various syntactic categories. For example, the boundaries marked out in the three lines presented above in (1), (2) and (3) are actually all coding 4 boundaries and they respectively represent the boundary between two coordinated NPís, between two sentences, and between the line-initial sentential adverb and the sentence it modifies. The following verse lines illustrate yet more possibilities of the syntactic constituents corresponding to the coding 4 boundary:

(34) *qing1 quan2* ↓ *shi2 shang4 liu2*
clear stream stone on flow
ëThe clear stream flows on the stonesí

(35) *tan4* ↓ *wei2 yao1 dai4 sheng4*
sigh tie waist belt extra
ë(I) sigh over the fact that my waist belt becomes longer (because Iím pining away)í

[12] These two demoting steps need to proceed in this sequential fashion with the demoting of derived 5ís preceding that of 4ís. The reason is that the demoting of 5ís results in yet more 4ís.

(36) ↓
xin1 zhi1 you1 yi3
heart prt worry interj
ëAh, my heart worriesí

The coding 4 boundaries in these three examples are respectively that between the subject NP and the VP, that between the V and the object NP, and that between the small clause and the interjection[13].

In some cases, syntax needs to look to the semantic interpretation of the line and the associated pragmatic considerations to produce a correct parsing of the line. This is illustrated in the following two cases. One is when a verse line contains no verbs and only juxtaposed NPís, as in (6) above, which is repeated below:

(i) (ii)
↓ ↓
(37) *gu3 dao4 xi1 feng1 shou4ma2*
ancient road west wind thin horse
ëThe ancient road, the west wind, and the thin horseí

To determine the relative strength of the two boundaries between the three NPís marked out above entails reference to the semantic interpretation of the line and certain pragmatic considerations. The line should be interpreted as ë*A thin horse (toils) on the ancient road in the west wind*í where the first two NPís describe the backdrop against which the referent of the third NP is embedded. Hence, the first two NPís are more closely connected to each other, and boundary (i) is weaker than boundary (ii); accordingly, boundary (ii) represents the strongest boundary within the line and is assigned coding 4.

A similar scenario is when the line contains more than one verb. Consider:

(i) (ii)
↓ ↓
(38) *feng1 hui2 lu4 zhuan3 bu2jian4 jun1*
mountain return road wind not see you
ëThe mountain reappears, and the road winds, and (I) cannot see you anymoreí

where the semantic interpretation of the line implies that boundary (ii) is stronger than boundary (i) and thus assigned coding 4.

The result of this trimming is that the only 5 is the underived one marking the boundary after the line-final syllable, and the only 4 is the one marking the strongest boundary in the line. All other 4ís and 5ís will be reduced to 3ís.

An illustration of the post-assignment operations necessitates that of the coding at the previous two stages, and the complete coding process is illustrated with the following verse line:

[13] Example (36) compellingly illustrates the relative nature of boundary strength: the boundary between V and its object NP is weak due to the theta relation, but here in the absence of any stronger boundary, it nonetheless constitutes the strongest break in the line and is therefore coded as 4.

(39) ↓

xiang1 *shu1* *he2* *chu4* *da2*
hometown letter whichplace arrive
ëWhere shall (my) letter home arrive?í

First, the pre-assignment stage trivially yields the coding pattern 33335.

Second, at the assignment stage, the following operations are executed. First, the coding 3 between ë*xiang1*í (hometown) and ë*shu1*í (letter) is reduced by one, because they are of the ëN+Ní structure where the first noun modifies the second[14]. Second, a similar reduction happens to the coding 3 between ë*he2*í (which) and ë*chu4*í (place). Third, the coding 3 between ë*shu1*í and ë*he2*í is increased by one because of the branching structure on both sides. Note that here as mentioned earlier, in case of branching as the alienating factor, the coding of a boundary is only increased by one no matter whether the branching occurs on one or both sides of it. Fourth, the coding 3 between ë*chu4*í and ë*da2*í is increased by one because of the branching structure of the verbal complement, i.e. ë*he2 chu4*í. Fifth, the coding 4 thus derived between ëchu4í and ëda2í is decreased by one because of the theta relation between the internal argument ë*he2 chu4*í (which place) and the verb ë*da2*í (arrive). Thus we have 24235.

It turns out that this coding pattern perfectly conforms to the coding profile template where coding 4 indeed marks the biggest break in the line, which is the boundary between the external argument and the predicate. As such, it need not undergo post-assignment.

For clarity sake, the complete coding process is illustrated below:

(40) *xiang1* *shu1* *he2* *chu4* *da2*
hometown letter whichplace arrive
ëWhere shall (my) letter home arrive?í

xiang1 *shu1* *he2* *chu4* *da2*
(i) Pre-assignment: 3 3 3 3 5
(ii) Assignment: Functor Branching Functor Branching
↓ ↓ ↓ ↓
2 4 2 4
Theta
↓
3
(iii) Post-assignment: None

→ Final coding: 24235

2.3.2 Exceptions regarding Coding 4

Now back to the post-assignment stage, as hinted earlier, two exceptions to the overall coding template mentioned above are permitted: first, for verse lines containing line-

[14] Alternatively, the reason might be that the constituent ë*xiang1 shu1*í is actually listed as a compound in the lexicon.

medial interjections, the coding necessarily contains two 4ís, and second, for certain two-syllabled lines, coding 4 may be legitimately absent.

In the former case, the boundaries on both sides of the interjection constitute the strongest boundary within the line, and as such are encoded as 4. This line type is most common in the second genre, *Chuci* and continues to encroach upon some earlier poems of the third genre, *Guti*, with ë*xi*í being the line-medial interjection. Two examples, respectively coded as 324425 and 24425, are as follows:

(41) *yu4 lan2 tang1* ***xi1*** *mu4 fang1*
bathe orchid water xi shower fragrance
ë(I) bathe myself in the orchid water and shower myself in fragranceí

(42) *wu3 yin1* ***xi1*** *fan2 hui4*
five sound xi exuberant luxurious
ëThe five sounds are exuberant and luxuriousí

The latter case only happens with certain two-syllabled lines, where the two syllables constitute a structure containing one of the binding factors identified above, for example, an NP of the ëmodifier + modifieeí structure, or a noun compound, as illustrated below:

(43) (i) *tuan2 shan4*
round fan
ëthe round faní

(ii) *guan3 xian2*
pipe string
ëthe pipe and string (music)í

In such lines, our practice is to allow for the absence of 4ís, and represent the tighter boundary between the two syllables by encoding the boundary strength as 25, which is exempt from subsequent post-assignment inspection.

Of course, 45 is still a possible coding type for two-syllabled lines, and predictably when these two syllables constitute a constituent containing no binding factor. For example, the following line is a ëN+Ví structure:

(44) *ren2 qiao1*
people quiet
ëPeople have quieted downí

2.4 Illustration of the coding scheme

To summarize, the coding scheme to encode the boundary strength of a verse line includes three stages: pre-assignment, assignment and post-assignment, with the algorithm for the assignment stage constituting the core. This algorithm works straightforwardly by identifying the formal grammatical factors bearing upon the boundary strength as binding or alienating. Methodologically, the coding scheme features an integration of bottom-up and top-down perspectives: at the pre- and post-assignment stages, the perspective is a top-down one whilst the bottom-up perspective is adopted at the assignment stage where the syntactic, semantic and lexical aspects of the line are considered.

We conclude this section by illustrating the application of the coding scheme with the following two examples.

(45) *huan2 qin3 meng4 jia1 qi1*
return sleep dream good time
ë(She) goes back to sleep, dreaming of good timesí

```
                          huan2   qin3    meng4  jia1    qi1
(i) Pre-assignment:           3      3        3     3      5
(ii) Assignment:           Theta            Theta  Functor
                              ↓                ↓      ↓
                              2                2      2
                                           Branching
                                               ↓
                                               3
(iii) Post-assignment:          Promotion
                                    ↓
                                    4
```

→ Final coding: 24325

(46) *bai2 tou2 gong1 nv3 zai4*
white head court lady is (there)
ëThe white-haired court lady is (still) thereí

```
                          bai2     tou2     gong1          nv3     zai4
(i) Pre-assignment:          3        3        3             3        5
(ii) Assignment:    Compounding  Functor  Compounding  Branching
                             ↓        ↓        ↓             ↓
                             2        2        2             4
                              Branching
                                  ↓
                                  3
(iii) Post-assignment: None
```

→ Final coding: 23245

Part II The ripe corpus

Shijing		
Line Type	**Number**	**Percentage**
2-syll	10	0.76%
3-syll	63	4.77%
4-syll	1134	85.91%
5-syll	70	5.30%
6-syll	29	2.20%
7-syll	11	0.83%
8-syll	3	0.23%
Total	1320	100%
Boundary Strength Coding Type	**Number**	**Percentage**
2-syll		
25	3	30%
35	3	30%
45	4	40%
Total	10	100%
3-syll		
145	6	9.52%
245	3	4.76%
345	3	4.76%
425	49	77.78%
435	2	3.17%
Total	63	100%
4-syll		
1345	1	0.09%
1415	2	0.18%
1425	52	4.59%
2345	43	3.79%
2415	66	5.82%
2425	459	40.48%
2435	39	3.44%
3145	1	0.09%
3245	106	9.35%
3445	3	0.26%
4235	19	1.68%
4315	14	1.23%
4325	329	29.01%
Total	1134	100%
5-syll		
23145	2	2.86%
23245	11	15.71%
23415	2	2.86%
24325	15	21.43%
24425	2	2.86%
32345	2	2.86%
32425	7	10%

33245	8	11.43%
42325	6	8.57%
42335	1	1.43%
43235	2	2.86%
43325	12	17.14%
Total	70	100%
6-syll		
232435	8	27.59%
233245	2	6.90%
242325	2	6.90%
243235	4	13.80%
243325	2	6.90%
333245	7	24.14%
432325	2	6.90%
433325	2	6.90%
Total	29	100%
7-syll		
2332435	3	27.27%
2343325	3	27.27%
3242415	1	9.09%
3243325	1	9.09%
3333245	3	27.27%
Total	11	100%
8-syll		
32343325	3	100%
Total	3	100%

Jiuge

Line Type	**Number**	**Percentage**
5-syll	49	19.76%
6-syll	128	50.59%
7-syll	73	28.85%
8-syll	1	0.40%
9-syll	2	0.79%
Total	253	100%
Boundary Strength Coding Type	**Number**	**Percentage**
5-syll		
14425	1	2.04%
24415	5	10.20%
24425	38	77.55%
24435	2	4.08%
34425	3	6.12%
Total	49	100%
6-syll		
134435	1	0.78%
234415	1	0.78%
234425	14	10.94%
314425	12	9.38%
314435	4	3.13%

324415	1	0.78%
324425	81	63.28%
324435	14	10.94%
Total	128	100%
7-syll		
2344235	4	5.48%
2344325	5	6.85%
2443325	1	1.37%
3144235	1	1.37%
3144315	4	5.48%
3144325	2	2.74%
3244235	1	1.37%
3244325	53	72.60%
3324425	2	2.74%
Total	73	100%
8-syll		
32442325	1	100%
Total	1	100%
9-syll		
324423325	1	50%
332443325	1	50%
Total	2	100%

Guti		
Line Type	**Number**	**Percentage**
4-syll	30	3.56%
5-syll	431	51.13%
6-syll	5	0.59%
7-syll	367	43.53%
8-syll	10	1.19%
Total	843	100%
Boundary Strength Coding Type	**Number**	**Percentage**
4-syll		
1425	1	3.33%
2415	4	13.33%
2425	10	33.33%
2435	2	6.67%
4315	2	6.67%
4325	11	36.67%
Total	30	100%
5-syll		
13245	1	0.23%
14235	12	2.78%
14315	1	0.23%
14325	20	4.64%
23425	2	0.46%
24235	55	12.76%
24315	12	2.78%
24325	238	55.22%

33245	1	0.23%
34425	3	0.70%
42335	1	0.23%
43235	58	13.46%
43315	2	0.46%
43325	25	5.80%
Total	431	100%
6-syll		
132445	1	20%
232445	2	40%
332445	2	40%
Total	5	100%
7-syll		
1234235	2	0.55%
1234325	3	0.82%
1314235	1	0.27%
1324135	1	0.27%
1324235	2	0.55%
1324325	4	1.09%
1423325	3	0.82%
1433235	1	0.27%
2314235	2	0.55%
2314325	3	0.82%
2324135	1	0.27%
2324235	34	9.26%
2324315	1	0.27%
2324325	88	23.98%
2334325	1	0.27%
2344235	4	1.09%
2344325	3	0.82%
2413235	2	0.55%
2413325	1	0.27%
2423135	7	1.91%
2423235	24	6.54%
2423315	2	0.55%
2423325	44	11.99%
2432325	1	0.27%
2433135	1	0.27%
2433235	18	4.90%
2433325	9	2.45%
3144315	1	0.27%
3244315	4	1.09%
3244325	13	3.54%
3314315	1	0.27%
3314325	1	0.27%
3324235	10	2.72%
3324315	1	0.27%
3324325	21	5.72%
3434325	1	0.27%
4312335	1	0.27%
4313325	3	0.82%

4323135	1	0.27%
4323235	12	3.27%
4323325	29	7.90%
4333235	5	1.36%
Total	367	100%
8-syll		
23244325	5	50%
32344235	1	10%
33244235	1	10%
33244325	3	30%
Total	10	100%

Jinti		
Line Type	**Number**	**Percentage**
5-syll	434	56.81%
7-syll	330	43.19%
Total	764	100%
Boundary Strength Coding Type	**Number**	**Percentage**
5-syll		
14235	10	2.30%
14315	1	0.23%
14325	2	0.46%
23245	5	1.15%
23435	2	0.46%
24115	1	0.23%
24135	3	0.69%
24235	96	22.12%
24315	7	1.61%
24325	228	52.53%
43135	2	0.46%
43235	62	14.29%
43325	15	3.46%
Total	434	100%
7-syll		
1234125	1	0.30%
1234235	4	1.21%
1234325	7	2.12%
1324235	6	1.82%
1324325	7	2.12%
1423325	6	1.82%
1433235	1	0.30%
1433325	1	0.30%
2314135	2	0.60%
2314235	4	1.21%
2314315	1	0.30%
2314325	3	0.91%
2324135	3	0.91%
2324235	44	13.33%
2324315	5	1.52%

2324325	78	23.64%
2334235	1	0.30%
2334325	2	0.60%
2412335	1	0.30%
2413235	2	0.60%
2423235	16	4.85%
2423315	1	0.30%
2423325	28	8.48%
2433125	1	0.30%
2433135	1	0.30%
2433235	30	9.09%
2433325	2	0.60%
3234235	1	0.30%
3234315	1	0.30%
3234325	1	0.30%
3243325	1	0.30%
3314235	1	0.30%
3314325	1	0.30%
3324235	5	1.52%
3324325	13	3.94%
4312335	1	0.30%
4313235	2	0.60%
4313325	1	0.30%
4323235	18	5.45%
4323315	2	0.60%
4323325	23	6.97%
4333235	1	0.30%
Total	330	100%

Ci		
Line Type	**Number**	**Percentage**
2-syll	10	1.33%
3-syll	125	16.60%
4-syll	261	34.66%
5-syll	128	17.00%
6-syll	112	14.87%
7-syll	111	14.74%
8-syll	2	0.27%
9-syll	4	0.53%
Total	753	100%
Boundary Strength Coding Type	**Number**	**Percentage**
2-syll		
15	2	20%
25	4	40%
35	2	20%
45	2	20%
Total	10	100%
3-syll		
245	48	38.40%

345	4	3.20%
415	7	5.60%
425	52	41.60%
435	14	1.12%
Total	125	100%
4-syll		
1425	6	2.30%
2345	8	3.07%
2415	8	3.07%
2425	143	54.79%
2435	25	9.58%
3245	5	1.92%
3435	3	1.15%
4315	4	1.53%
4325	57	21.84%
4335	2	0.77%
Total	261	100%
5-syll		
13245	1	0.78%
14325	1	0.78%
23345	2	1.56%
23425	1	0.78%
24135	2	1.56%
24235	21	16.41%
24315	2	1.56%
24325	39	30.47%
32425	1	0.78%
34235	4	3.13%
41325	5	3.91%
42325	23	17.97%
42335	3	2.34%
43135	1	0.78%
43235	9	7.03%
43315	1	0.78%
43325	12	9.38%
Total	128	100%
6-syll		
132435	1	0.89%
142325	5	4.46%
142335	1	0.89%
143235	1	0.89%
231435	2	1.79%
232345	1	0.89%
232415	1	0.89%
232425	5	4.46%
232435	2	1.79%
241315	1	0.89%
241325	3	2.68%
242315	1	0.89%
242325	32	28.57%
242335	7	6.25%

243315	2	1.79%
243325	18	16.07%
332425	2	1.79%
342325	2	1.79%
423235	1	0.89%
432315	1	0.89%
432325	20	17.86%
432335	2	1.79%
433325	1	0.89%
Total	112	100%
7-syll		
1314325	1	0.90%
1324235	2	1.80%
1324325	3	2.70%
1334325	1	0.90%
1423325	1	0.90%
1433235	1	0.90%
2314235	2	1.80%
2314325	2	1.80%
2324235	17	15.32%
2324325	20	18.02%
2324335	2	1.80%
2334315	1	0.90%
2334335	1	0.90%
2343325	1	0.90%
2413235	1	0.90%
2413325	1	0.90%
2423135	2	1.80%
2423235	5	4.50%
2423325	10	9.01%
2432325	1	0.90%
2433235	11	9.91%
2433325	4	3.60%
3244325	1	0.90%
3324325	9	8.11%
4313235	1	0.90%
4323235	1	0.90%
4323315	1	0.90%
4323325	6	5.41%
4333235	2	1.80%
Total	111	100%
8-syll		
43232325	1	50%
43323325	1	50%
Total	2	100%
9-syll		
243323235	1	25%
431323235	1	25%
432323235	2	50%
Total	4	100%